基尼系数的算法

■徐万坪

基尼系数的算法

1905 年，统计学家马克斯·洛伦兹(Max·Lorrenz)首先提出了一个用以描述收入或财富分配不均等程度的曲线，即洛伦兹曲线。洛伦兹曲线是指按照人均收入水平由低到高排序，横轴为人口累计比率 X，纵轴为收入累计比率 Y 的平面直角坐标体系下的一条曲线 Y=Y(X)。当洛伦兹曲线是 Y=X 时，它表示人均占有社会财富或人均收入水平的绝对平均；当洛伦兹曲线是横轴与 X=100 处的垂线所构成的直角折线时，它表示收入分配的绝对不平均。一般来说，洛伦兹曲线是位于上述两种极端情况中间的一条曲线 Y=Y(X)，该曲线与对角线越接近，收入分配越平均；反之，收入分配越不平均。(见图 1)。

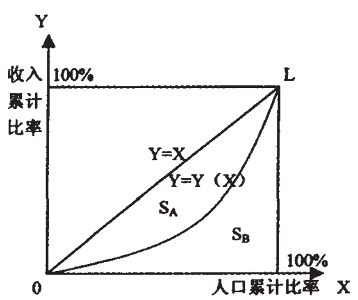

图 1

1912 年，意大利经济学家科拉多·基尼(Corrado·Gini)又以 Lorrenz 曲线的图形为基础，提出了基尼系数，基尼系数的含义是指在全部居民收入中，用于进行不平均分配的那部分收入占总收入的百分比。它是测度收入分配不平等程度的指标，是在某一段时期内从总体上对一定范围内居民的收入分配的不平等程度的描述。基尼系数是被人们普遍接受的并行之有效的方法，作为联合国规定的社会经济发展指标之一。基尼系数 $G=S_A/(S_A+S_B)$，其中 S_A 是指 45 度线 Y=X 与洛伦兹曲线 Y=Y(X)所围成的面积，而 SB 是洛伦兹曲线 Y=Y(X)与 OX、XL 所围成的面积。由于 S_A+S_B 等于 1，所以 $G=2S_A$。若知道洛伦兹曲线，则基尼系数 $G=\int_0^1[X-Y(X)dX]/\int_0^1 XdX=1-2\int_0^1 Y(X)dX$。国内外学者提出了许多基尼系数的计算方法，包括直接计算方法、面积法、拟合曲线法等等，但是缺乏一个比较系统的总结。本文对基尼系数的一些常用公式进行了推导和介绍，并提出一些在实际应用公式时容易忽略和混淆的问题。

基尼系数的计算方法

(一)直接计算法

基尼在 1912 年给出了基尼系数的计算公式，这种算法并不依赖于洛伦兹曲线。

$$G=\sum_{i=1}^{n}\sum_{j=1}^{n}\left|Y_i-Y_j\right|/2n^2\mu$$

它表示按人口分布所形成的收入差距对收入总体期望偏离的相对程度。其中，表示收入差的绝对值，表示收入的期望值。表示收入的平均差，而将收入的平均差除以收入期望值，得出收入平均差偏离收入期望值的期望程度，也就是基尼系数。该种方法的计算量较大，可以通过计算机操作。若将直接计算法应用于分组收入中，也就是《中国统计年鉴》所提供的数据，因为该法必须估算出每组的平均收入，如果组数比较少的话，会大大降低其计算的准确性。而且应用于分组收入的计算时，还必须进行各种假设：如各组人口分布对应的收入均值等于各人口组中间值对应的收入数值；每一组别的收入分布随着人口累计呈等比例增加的规律变化等。(钱敏泽，2002)。如果按照采集数据的种类来分类，基尼系数的方法可以大略可分为按个体收入来计算的方法，以及按分组收入来计算的方法。由于前种方法的数据不容易得到，官方公布的数据也大都是分组收入，若要应用前种方法，必须得到个体收入，或者将分组收入换算成个体收入，其中隐含了许多假设，使得计算的基尼系数不准确，所以在实际应用中计算基尼系数的公式以后一种方法为主。

(二)面积法

1.上梯形面积法

设有 n 个分组，各组的收入总量为 Y_i，各组收入占总体收入的比率为 $y_i=Y_i/\Sigma Y_i$，以各组人口数 P_i 来表示各组人口规模的不同，各组人口占总体人口的比率为 $p_i=P_i/\Sigma P_i$，各组人均收入水平 $Z_i=Y_i/P_i$ $(I=1,2,\cdots n)$ 且 $Z_1<Z_2<Z_3<\cdots<Z_n$。如图 2 阴影部分表示，SA 的面积等于 n 个矩形面积之和加上 n 个三角形面积之和再减 1/2。

注：本文系社科基金项目(03BjL038)的阶段性成果。

度的选取。进一步精确而言，标准概率术语的引用，意味着套利价格取决于主观概率测度的支撑，但是不随同一等值概率测度集合中的特定概率测度的选取而改变。金融术语中，上述论断可被重述为：所有投资者在原生证券的期货价格波动幅度上的认识达成一致；尽管他们在相应的主观概率评估上有所不同。

2、期货市场鞅测度

寻求一概率测度 p，使得其期货价格过程(未贴现)遵循 p 鞅分布。一概率 p，如果存在，将由下列等式决定：

$$f_0=E_{p}(f_T)=p\,f^u+(1-p)f^d$$

容易得出

$$p\{\omega_1\}=p=\frac{f_0-f^d}{f^u-f^d},\quad p\{\omega_2\}=1-p=\frac{f^u-f_0}{f^u-f^d}$$

这表明鞅方法同样可用于期货市场，只要对鞅测度概念稍加修改。

借助数理金融传统术语，人们可能得出结论：风险中性经济体系具有期货价格过程公平博弈性质。应该记住：风险中性即期经济中贴现股票价格(与股票价格本身相对)可将公平博弈模型化。

(作者单位/中国人民大学财政金融学院)

(责任编辑/李友平)

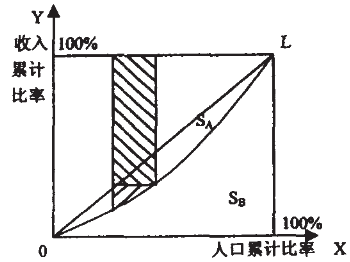

图2

$$\frac{1}{2}+S_A=\frac{1}{2}p_1y_1+p_1(1-y_1)+\frac{1}{2}(p_2-p_1)(y_2-y_1)+(p_2-p_1)(1-y_2)+\frac{1}{2}(p_3-p_2)(y_3-y_2)+(p_3-p_2)(1-y_3)+\cdots+\frac{1}{2}(p_n-p_{n-1})(y_n-y_{n-1})+(p_n-p_{n-1})(1-y_n)$$

因为 $p_n=y_n=1$，所以对上式展开化简可以得到

$$G=2S_A(p_iy_{i+1}-p_{i+1}y_i)+2p_n-p_ny_n-1$$

$$=\sum_{i=1}^{n}(p_iy_{i+1}-p_{i+1}y_i)$$

若把全部人口按收入的从小到大排列，并等分为 k 组，设第 i 组的收入额为 Y_i，则该组的收入额占全部收入的比重为 $y_i=Y_i/\Sigma Y_i$，(i=1, 2, …, k)，且有 $y_1\leqslant y_2\cdots\leqslant y_k$，$y_1+y_2+\cdots+y_k=1$。从图2可以看出，$S_A$ 的面积仍然是等于 n 个矩形面积之和加上 n 个三角形面积之和再减 1/2。

$$\frac{1}{2}+S_A=\frac{1}{k}\left(\frac{y_1}{2}+y_2+\cdots+y_k\right)+\frac{1}{k}\left(\frac{y_2}{2}+y_3+\cdots+y_k\right)+\cdots+\frac{1}{k}\times\frac{y_k}{2}$$

$$=\frac{1}{k}(y_1+2y_2+\cdots+ky_k)-\frac{1}{2k}(y_1+y_2+\cdots+y_k)$$

$$S_A=\frac{1}{k}(y_1+2y_2+\cdots+ky_k)-\left(\frac{1}{2k}+\frac{1}{k}\right)$$

$$G=2S_A=\frac{2}{k}(y_1+2y_2+\cdots+ky_k)-\frac{K+1}{k}$$

$$=\frac{2}{k}\sum_{i=1}^{K}iy_i-\frac{K+1}{k}$$

2.下梯形面积法

如图4阴影部分所示，可以先计算出 S_B 的面积，等于 n 个梯形之和，然后通过 S_B 求出 S_A 的值，最后求得基尼系数。

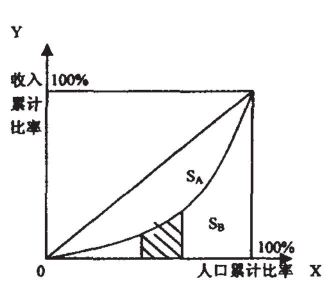

图4

$$S_B=\frac{1}{2}\times y_1\times p_1+\frac{1}{2}\times(y_1+y_2)\times(p_2-p_1)+\cdots+\frac{1}{2}\times(y_{n-1}+y_n)\times(p_n-p_{n-1})$$

$$\frac{1}{2}\times y_1\times p_1+\frac{1}{2}\times\sum_{i=1}^{n-1}(y_i+y_{i+1})\times(p_{i+1}-p_i)$$

$$G=\frac{1/2-S_B}{1/2}=1-2S_B$$

$$=1-y_1p_1-\sum_{i=1}^{n-1}(y_i+y_{i+1})\times(p_{i+1}-p_i)$$

当所采集的样本数据包含个体的收入时，可以采用以下公式，仍然是用下梯形的面积来求基尼系数。

$$S_A=\left(\frac{1}{2}\times\frac{1}{n}\times\frac{1}{n}-\frac{1}{2}\times\frac{1}{n}\times\frac{Y_1}{n\mu}\right)+\left[\frac{1}{2}\times\frac{1}{n}\times\left(\frac{1}{n}+\frac{2}{n}\right)-\frac{1}{2}\times\frac{1}{n}\times\left(\frac{Y_1}{n\mu}+\frac{Y_1+Y_2}{n\mu}\right)\right]+\cdots+\left[\frac{1}{2}\times\frac{1}{n}\times\left(\frac{1}{n}+\frac{2}{n}+\cdots+\frac{n}{n}\right)-\frac{1}{2}\times\frac{1}{n}\times\left(\frac{Y_1}{n\mu}+\frac{Y_1+Y_2}{n\mu}+\cdots+\frac{Y_1+Y_2+\cdots+Y_n}{n\mu}\right)\right]$$

$$G=2S_A=\frac{2}{n}\sum_{i=1}^{n}\left(\frac{i}{n}-\frac{1}{n\mu}\sum_{j=1}^{i}Y_j\right)$$

下面的公式也比较常用：

$$G=2S_A=\frac{2}{n}\sum_{i=1}^{n}\left(\frac{i}{n}-\frac{1}{n\mu}\sum_{j=1}^{i}\right)Y_j$$

$$=\frac{2}{n}\times\left(\frac{1}{n}+\frac{2}{n}+\cdots+\frac{n}{n}\right)-\frac{1}{n}\times\frac{1}{n\mu}[Y_1+(Y_1+Y_2)+\cdots+(Y_1+Y+\cdots+Y_n)]$$

$$=\frac{n+1}{n}-\frac{2}{n^2\mu}\sum_{i=1}^{n}(n-i+1)Y_i$$

3.矩形面积法

如图3阴影部分所示，S_A 的面积等于 n 个矩形面积之和减 n 个三角形面积之和再减 1/2，可以计算出 S_A 的面积，进而计算出基尼系数(白雪梅，赵松山，)。

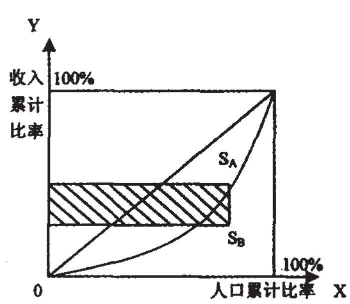

图3

$$S_A=\sum_{i=1}^{n}\left(y_i\times\sum_{j=1}^{n}p_j\right)-\frac{1}{2}\times\sum_{i=1}^{n}y_i\times p_i-\frac{1}{2}$$

$$G=2S_A=2\sum_{i=1}^{n}\left(y_i\times\sum_{j=1}^{n}p_j\right)-\sum_{i=1}^{n}y_i\times p_i-1$$

当 $p_1=p_2=p_3=\cdots=p_n=1/n$ 时，

$$G=\frac{2}{n}\sum_{i=1}^{n}i\times p_i-\frac{n+1}{n}$$

(三)拟合曲线法

1.幂函数

可以用幂函数 $Y=\alpha X\beta$ 来拟合洛伦兹曲线 $Y=Y(X)$，则

$$G=G=\int_0^1[X-Y(X)dX]/\int_0^1 XdX=\int_0^1(X-\alpha X^{\beta})dX/1/2=1-2\alpha/(1+\beta)$$

通过抽样调查的统计资料估计出参数 α 和 β，就可以根据上述公式计算出基尼系数 G 的值。

2.其他曲线

假设洛伦兹曲线为 $y=[1-(1-x)^{\alpha}]^{\frac{1}{\beta}}$，$0\leqslant\alpha,\beta\leqslant1$。为求出 G 值，我们利用收入分组数据求出假设洛伦兹曲线 $y=[1-(1-x)^{\alpha}]^{\frac{1}{\beta}}$ 中 α，β 的估计值。为此，我们令 $1-x=u$，则有 $u^{\alpha}+y^{\beta}=1$，然后用非线性回归的方法求出的估计值，于是

$$G=1-2S_B=1-2\int_0^1[1-(1-x)^{\varepsilon}]_{\frac{1}{\beta}}dx$$

$$=1-\frac{2}{\alpha}B\left(\frac{1}{\alpha},\frac{1}{\beta}+1\right)$$

$$=1-\frac{1}{\alpha\beta}\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\beta}\right)/\Gamma\left(\frac{1}{\alpha}+\frac{1}{\beta}+1\right)$$

$$=1-2\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\beta}\right)/\left[\alpha+\beta\Gamma\left(\frac{1}{\alpha}+\frac{1}{\beta}\right)\right]$$

结论

上面总结了基尼系数的计算公式，在计算时应注意以下两个常出现的问题：一是要区别基尼系数计算公式中收入 Y 和累计收入比率 y，采用不同的变量，其计算公式是不一样的，由于一般的数据采集包括《中国统计年鉴》均是给出分组收入，因此公式用累计收入比率 y 较为便捷，而统计中有时缺乏个人的收入 Y，所以包含收入 Y 的计算公式在实际计算中常被忽略，但对包含收入 Y 的计算公式有一个全面的了解是很有必要的。二是要区别每户收入和个人收入，采用不同类型的公式，计算出的基尼系数也是不同的。《中国统计年鉴》所提供的收入是每户收入，所以若要准确计算个人收入的基尼系数，必须把每户收入换算成为个人收入。在实际计算基尼系数时，根据所获取的样本数据类型，选取适合的计算公式是很必要的，这样能够较为准确地计算出基尼系数，同时可考虑其他计算方法得出的结果，为用基尼系数分析收入分配状况提供更好的定量依据。

(作者单位/厦门大学财政系)

(责任编辑/亦　民)